

AD-A176 350

NAMRL/NADC JOINT REPORT
NAMRL SPECIAL REPORT 86-3
NADC REPORT 86105-60 (VOL. 3)

THEORETICAL DEVELOPMENT FOR IDENTIFYING
UNDERLYING INTERNAL PROCESSES

VOLUME 3

RANDOM SAMPLING OF DOMAIN VARIANCES:
A NEW EXPERIMENTAL METHODOLOGY

R. J. Wherry, Jr.

JOINT REPORT

DTIC
ELECTE
FEB 04 1987
S D

Naval Aerospace Medical Research Laboratory
Pensacola, Florida

Naval Air Development Center
Warminster, Pennsylvania



September 1986

DTIC FILE COPY

Approved for public release, distribution unlimited.

82- 2 4 019

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified			1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.		
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE					
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NAMRL-SPECIAL REPORT 86-3 NADC REPORT 86105-60 (Vol.3)			5. MONITORING ORGANIZATION REPORT NUMBER(S)		
6a. NAME OF PERFORMING ORGANIZATION Analytics, Inc.		6b. OFFICE SYMBOL (If applicable)		7a. NAME OF MONITORING ORGANIZATION (Joint Monitoring) *Naval Aerospace Medical Research Laboratory **Naval Air Development Center	
6c. ADDRESS (City, State, and ZIP Code) 2500 Maryland Road Willow Grove, Pennsylvania 19090				7b. ADDRESS (City, State, and ZIP Code) *Naval Air Station, Pensacola, FL 32508-5700 **Warminster, PA 18974-5000	
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Naval Air Systems Command		8b. OFFICE SYMBOL (If applicable) (Code 330J)		9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER NAVAIRSYSCOM 61153N WR04210001.6142 NADC N62269-82-D-0131	
8c. ADDRESS (City, State, and ZIP Code) Washington, DC 20361				10. SOURCE OF FUNDING NUMBERS	
				PROGRAM ELEMENT NO. 61153N	PROJECT NO. WR04210001
11. TITLE (Include Security Classification) THEORETICAL DEVELOPMENTS FOR IDENTIFICATION UNDERLYING INTERNAL PROCESSES. VOLUME 3: RANDOM SAMPLING OF DOMAIN VARIANCE: A NEW EXPERIMENTAL METHODOLOGY					
12. PERSONAL AUTHOR(S) R. J. Wherry, Jr.					
13a. TYPE OF REPORT Interim		13b. TIME COVERED FROM TO 8/86		14. DATE OF REPORT (Year, Month, Day) 86-9	
15. PAGE COUNT 70					
16. SUPPLEMENTARY NOTATION					
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) Random Sampling, Domain Variance, Human Factors Engineering		
FIELD	GROUP	SUB-GROUP			
19. ABSTRACT (Continue on reverse if necessary and identify by block number) Modern aviation weapon systems impose increasingly complex and highly demanding command/control and information processing requirements on aircrew personnel. Improved assessment methods and more complete knowledge of human performance capabilities and limitations in high-demand, multi-task environments are needed to better match operator to the changing human roles in emerging aviation systems. The human engineering and human performance assessment and prediction technologies have, unfortunately, failed to keep pace with increasingly sophisticated airborne weapons systems currently being developed. The paucity of scientifically-based knowledge concerning the underlying human perceptual, cognitive, and motor processes makes it impossible to confidently influence system design or to be able to predict human and/or system performance in complex situations. This lack of knowledge stems primarily from not having firmly established:					
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS				21. ABSTRACT SECURITY CLASSIFICATION Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL J. O. Houghton, CAPT MC USN				22b. TELEPHONE (Include Area Code) (0041) 452-3286	
				22c. OFFICE SYMBOL Code 00	

Security classification of this page: UNCLASSIFIED

ABSTRACT (Continued)

(a) the numbers of, the nature of, the underlying internal processes, (b) the distributions of time and accuracy capabilities for those processes, (c) the extent to which individual differences among those processes are stable across tasks which use those processes, (d) the nature or identification of task factors which cause (or accompany) the invoking of some processes but not others, and (e) possible fatigue, recovery, and/or interference in internal processing brought about by repeated and/or competing demands on those processes.

Resolution of these basic problems is seen as central in elevating both human engineering design/evaluation and human performance assessment/prediction technologies to a more responsive level for the Navy's RDT&E system acquisition process and for meeting the Navy's personnel selection, assignment, and training requirements.

The Theory of Underlying Internal Processes (UIPs) was described in the first volume of this series. In that document, it is shown that a factor analysis of the correlations (across subjects) of the response times for a battery of tasks should lead to the discovery of independent factors which represent the underlying processes which are common to two or more tasks in the battery, and that loadings on those factors must all be zero or positive. Such a factor structure is referred to as one having "positive manifold." This volume describes the modifications to the Hierarchical Factor Analysis (HFA) needed to arrive at such structures. This volume contains a discussion of a new experimental methodology which permits the simultaneous investigation of a large number of experimental variables each of which may have many possible levels. The need for such a methodology stems from the complex nature of tasks and environmental situations of interest to the Navy. The new technique is entitled Random Sampling of Domain Variance (RSDV) and is seen as alternative to the Analysis of Variance (ANOVA) method which has not been particularly successful in producing results which generalize to the real world.

Development and elucidation of the Random Sampling of Domain Variance experimental methodology has been accomplished under the Navy's special focus program entitled **Augmentation of Human Factors Technology Efforts** which has been jointly sponsored by the Engineering Psychology Programs of the Office of Naval Research and the Naval Air Systems Command.

Approved for public release; distribution unlimited.

THEORETICAL DEVELOPMENT FOR IDENTIFYING
UNDERLYING INTERNAL PROCESSES.

VOLUME 3

RANDOM SAMPLING OF DOMAIN VARIANCE:
A NEW EXPERIMENTAL METHODOLOGY

R. J. Wherry, Jr.*

Joint Report

Naval Aerospace Medical Research Laboratory
Naval Air Development Center
NAVAIRSYSCOM 61153N WRO4210001.6142

*Analytics, Inc.
Contract No. N62269-82-D-0131



Approved and Released by

Captain J. O. Houghton, MC USN
Commanding Officer
Naval Aerospace Medical
Research Laboratory

Approved and Released by

Captain E. J. Sturm
Commanding Officer
Naval Air Development Center



September 1986

By _____	
Distribution / _____	
Availability Codes	
Dist	Avail and/or Special
A-1	

SUMMARY PAGE

THE PROBLEM

Modern aviation weapon systems impose increasingly complex and highly demanding command/control and information processing requirements on aircrew personnel. Improved assessment methods and more complete knowledge of human performance capabilities and limitations in high-demand, multi-task environments are needed to better match operator to the changing human roles in emerging aviation systems. The human engineering and human performance assessment and prediction technologies have, unfortunately, failed to keep pace with increasingly sophisticated airborne weapons systems currently being developed.

The paucity of scientifically-based knowledge concerning the underlying human perceptual, cognitive, and motor processes makes it impossible to confidently influence system design or to be able to predict human and/or system performance in complex situations. This lack of knowledge stems primarily from not having firmly established: (a) the numbers of, the nature of, the underlying internal processes, (b) the distributions of time and accuracy capabilities for those processes, (c) the extent to which individual differences among those processes are stable across tasks which use those processes, (d) the nature or identification of task factors which cause (or accompany) the invoking of some processes but not others, and (e) possible fatigue, recovery, and/or interference in internal processing brought about by repeated and/or competing demands on those processes.

Resolution of these basic problems is seen as central in elevating both human engineering design/evaluation and human performance assessment/prediction technologies to a more responsive level for the Navy's RDT&E system acquisition process and for meeting the Navy's personnel selection, assignment, and training requirements.

The Theory of Underlying Internal Processes (UIPs) was described in the first volume of this series. In that document, it is shown that a factor analysis of the correlations (across subjects) of the response times for a battery of tasks should lead to the discovery of independent factors which represent the underlying processes which are common to two or more tasks in the battery, and that loadings on those factors must all be zero or positive. Such a factor structure is referred to as one having "positive manifold." This volume describes the modifications to the Hierarchical Factor Analysis (HFA) needed to arrive at such structures. This volume contains a discussion of a new experimental methodology which permits the simultaneous investigation of a large number of experimental variables each of which may have many possible levels. The need for such a methodology stems from the complex nature of tasks and environmental situations of interest to the Navy. The new technique is entitled Random Sampling of Domain Variance (RSDV) and is seen as alternative to the Analysis of Variance (ANOVA) method which has not been particularly successful in producing results which generalize to the real world.

In addition to the Theory of Underlying Internal Processes (presented in volume one), the development of the method for obtaining positive manifold factor structures (presented in volume two), and the Random Sampling of Domain Variance (RSDV) method described in this volume, two

other significant methodological developments have arisen during this project and are discussed in detail in the other volumes of this series:

Volume 4 - Task Domains of Naval Flight Officers (NFOs).

Volume 5 - Special Computer Applications in UIP/RSDV Studies.

Acknowledgments

Development and elucidation of the Random Sampling of Domain Variance experimental methodology has been accomplished under the Navy's special focus program entitled **Augmentation of Human Factors Technology Efforts** which has been jointly sponsored by the Engineering Psychology Programs of the Office of Naval Research and the Naval Air Systems Command. I am indebted to Mr. G. Malecki and Commander T. Jones, from those respective commands, for their support for this effort.

TABLE OF CONTENTS

1.	INTRODUCTION	
2.	BACKGROUND	
2.1	Dissatisfaction with Experimental Results	2-1
2.2	A Brief History of ANOVA	2-2
2.3	Fixed, Random, and Mixed Effects Models in ANOVA	2-4
2.4	The Practical Limitations of ANOVA	2-6
2.5	The Centrality of the Pearson Equation	2-8
2.6	The Role of Experimentation in Research	2-9
3.	DECISIONS EFFECTING HUMAN PERFORMANCE STUDIES	
3.1	Elements of the Situation to be Studied	3-1
3.1.1	Definition of a Domain	3-2
3.1.2	Problems with ANOVA Studies	3-3
3.1.3	Further Difficulties with Field Studies	3-5
3.1.4	The Need for "Domain Specification"	3-6
3.1.5	Inability to Resolve Unexpected Data Cases	3-7
3.2	Measures of Performance to be Collected	3-7
3.3	Methods of Data Analysis	3-8
3.3.1	Univariate Measures and Methods	3-8
3.3.2	Bivariate Measures and Methods	3-11
3.3.3	Multivariate Methods: Multiple Correlation	3-12
3.3.4	Multivariate Methods: Factor Analysis	3-15
4.	PRINCIPLES AND APPLICATIONS OF RANDOM SAMPLING	
4.1	Random Sampling Defined	4-1
4.2	"Randomization" in Experiments	4-2
4.3	Statistics Based on Random Samples from a Population	4-3
4.4	The Statistics of Multiple Random Samples	4-4
4.5	Random Sampling of Items with Multiple Attributes	4-4
4.6	Random Sampling in Simulation and Models	4-5
4.7	The Wherry, Jr., Simulated Data Generation Technique	4-6
4.7.1	Generating Simulated Rating Data	4-7
4.7.2	The Method for Generating the Simulated Data	4-8
4.7.3	Other Applications of the Simulated Data Generation Technique	4-9

TABLE OF CONTENTS (continued)

5.	THE RANDOM SAMPLING OF DOMAIN VARIANCE (RSDV) TECHNIQUE	
5.1	The Objective of the RSDV Technique	5-1
5.2	The Efficacy of Random Sampling	5-2
5.3	Procedures for Conducting RSDV Studies	5-2
	5.3.1 Specifying the Domains of Interest	5-3
	5.3.2 Selecting the Situations to be Studied	5-9
6.	SUMMARY	
	REFERENCES	

1. INTRODUCTION

The Random Sampling of Domain Variance (RSDV) concept presented here represents a new approach for experimental design and analysis. Much dissatisfaction has been expressed for many years over the inability of results derived from Analysis of Variance (ANOVA) studies to be applied to practical problems of the real world. These problems are well documented and will be discussed in subsequent sections along with how ANOVA, itself, contributes to those problems. The advantages and disadvantages of both traditional experimental techniques and field studies are compared, and both are found wanting. The RSDV concept represents a novel and rationally defensible approach to obtaining results, based on data collected in laboratory settings, which should more validly generalize to the real world and, thus, prove valuable to psychologists of all walks, regardless of whether they have a theoretical or practical bent. The RSDV concept also appears to furnish an excellent bridge to link laboratory and field studies.

The major feature of the RSDV concept involves the extensive usage of random sampling theory, on which virtually all tests of significance are based. Because of this, the RSDV concept represents a natural progression in statistical methods, and one which, hopefully, will be readily grasped by most behavioral scientists and practitioners. As with any new method, standard and convenient procedures need to be developed so that RSDV users can report their studies in formats easily understood by others in the field. While no pretense is made that this type of effort has, in any way, been completed, at least some effort has been made toward this goal. Some initial suggestions have been provided for researchers on the activities required in conducting RSDV studies and how investigators might report both these activities and the findings they obtain from such studies.

2. BACKGROUND

2.1 DISSATISFACTION WITH EXPERIMENTAL RESULTS

Psychology has long suffered from an enormous gulf between the accrual of experimental results collected in laboratory settings and the applicability of those same results to practical situations of the real world. Nearly a decade ago, Simon (1975) documented the widespread disappointment, dissatisfaction, and disillusionment with the practical value of results from laboratory experimentation in all fields of psychology. His paper reviewed nearly 240 analysis of variance (ANOVA) studies from 118 articles published in the journal Human Factors during the fourteen years between 1958 and 1972. Among his findings were that over 92 percent of those studies investigated three or fewer experimental variables. On the average, these variables accounted for only 45 percent of the total variance. Over 98 percent of these studies investigated four or fewer variables and accounted for only 61 percent of the variance.

Dunnette (1966), in his review of four American Psychological Association (APA) journals, had found similar results nearly a decade earlier. It should be clear to most psychologists that human performance in the complex, real world is governed by far more than three or four variables. Thus, the results which Simon and Dunnette found certainly should not astonish us. Indeed, we would be more surprised if so few variables were to be responsible for so much variance in human performance in real world situations. Simon, however, pointed out that experimenters, too often, artificially enhanced the proportion of variance accounted for by using averaged rather than actual scores in their analyses. To this criticism of ANOVA users, it should also be mentioned that, often, only two highly divergent levels of an experimental variable are investigated. This practice, by removing the central portion of the possible effects, will also result in significant overestimates of real world variance that

is actually explainable. For both of these reasons, the identical variables (which apparently work so well in laboratory settings) can be expected to account for far less variance in the real world. This phenomena, which has been experienced all too frequently, has lead, in part, to the disappointment, dissatisfaction, and disillusionment mentioned earlier.

The widespread acceptance and advocacy of ANOVA, a technique championed for so long by many psychologists as the "preferred" experimental design and analysis technique, may be, perhaps, the real culprit. The ready availability of ANOVA, with its statistical defensible rationale, initially offered experimental psychology the scientific respectability it originally needed and which it had been seeking. At the same time, however, it must also be admitted that the wholesale embracing of ANOVA, with its lack of ability to simultaneously investigate many different levels of a large number of experimental variables, ultimately has delayed psychology from becoming the scientific discipline capable of predicting and/or explaining the behavior of humans confronted with the complex situations of the real world.

2.2 A BRIEF HISTORY OF ANOVA

It is desirable to he briefly review the fifty-year history of the analysis of variance technique and how it came to play such a dominant role in experimental psychology. ANOVA, as many now know, is a technique borrowed from procedures originally developed for agricultural research. R. A. Fisher (1934), the leading proponent and early populariser of ANOVA, pointed out that ANOVA's *"one claim to attention lies in its convenience."* First, it provided a convenient procedure for summarizing a *"mass of statistical data,"* a task which was far more difficult in the 1930s when data analysis calculations were manually intensive and fraught with opportunities for making errors. Secondly, Fisher pointed out that ANOVA was *"convenient in facilitating and reducing to a common form all the tests of significance which we may want to apply."*

Fisher's major contributions were in the area of tests of significance, and it is certainly understandable that he might favor a procedure that emphasized this aspect of statistics even if it neglected reporting the relationships between the experimental variables and the criterion. What Fisher did was to greatly emphasize the issue of statistical significance over that of practical significance. One cannot help but wonder if Karl Pearson, who had given correlation theory its mathematical foundations, would not have selected a more balanced presentation of both the statistical and practical results of studies. But by 1933, Pearson had retired from the scene. Fisher had succeeded him to the prestigious chair of the Galton Professorship at University College in London and had, thus, inherited the most influential position in statistics at that time. If Fisher advocated ANOVA, it must be good! It is worth mentioning that, while ANOVA represented a new methodology by which to accomplish various calculations and a new way to present and summarize one's results, it neither invented the concept of tests of significance, nor created any new tests. Instead, it adopted those which had been completely worked out in the early 1920s.

While ANOVA gained some early advocates and disciples, its usefulness was certainly being questioned by other leaders in statistics. Peters and Van Voorhis (1940), for example, in a faintly disguised allusion to ANOVA's origin in agricultural research, stated, "*It is always pedantic to make forced use of statistical devices borrowed from another field when they only poorly fit.*" Peters and Van Voorhis made a large distinction between the type of research which could make use of ANOVA and what they considered to be "positive" or "constructive" research. They acknowledged that ANOVA could be used as an initial step to make a "*rough, preliminary test*" of a hypothesis before "*going to the expense of the elaborate setup needed for a thorough investigation. ... But for the positive side of research, the investigator will need the standard procedures of classical statistics, such as correlation, curve fitting, and contrast of correlated matched groups. Constructive research is just ready to begin where analysis of variance leaves off.*" To conclude that such sentiments irritated the many advocates of ANOVA would be an understatement. Such comments were but the earliest exchanges of a

continuing disagreement between the proponents of the ANOVA approach and the advocates of classical correlational approaches. Even today, there are still remnants of misunderstanding on both sides of this issue.

Despite such dire warnings and reservations, the convenience of the ANOVA procedure did find increasingly rapid acceptance among many psychologists who preferred to investigate hypotheses about the effect of one or, perhaps, two variables on some criterion of performance. Studies which investigated only one, two, or three variables were (and still are) relatively easy and inexpensive to conduct, and ANOVA offered a convenient procedure and an acceptable way of reporting research findings that made them readily publishable. Soon, both courses and textbooks appeared which, in a large measure, helped to institutionalize the usage of ANOVA among many aspiring psychologists.

Among the well known authors of those early books favoring ANOVA were Cochran and Cox (1950, 1957). In the preface to the 1957 edition of their book on experimental designs, they discussed the growing usage of ANOVA at that time. They state, *"Another encouraging trend is that workers in these areas, although still willing to utilize appropriate designs taken from agricultural experimentation, have begun to examine their own problems of experimentation and to produce new designs better suited to their particular conditions."* It is of some interest to note their acknowledgment of the predominance of univariate studies at that time when they stated, *"The recent literature also reflects a move toward greater depth and comprehensiveness in experimental work, as instanced by numerous papers devoted to experimentation with more than one factor."*

2.3 FIXED, RANDOM, AND MIXED EFFECTS MODELS IN ANOVA

No ANOVA discussion, however brief, should neglect the distinction between "fixed," "random," and "mixed" effects models employed by that technique. Other sources can provide far more extensive discussions of this topic, so only the briefest of treatments of this subject will be presented here. The basic distinction among the various models centers on how one decides which levels of a given experimental variable will be used in an experiment. If the researcher arbitrarily

makes this decision, regardless of how rational or irrational the reasons may be, then the levels of that experimental variable are deemed to be "fixed" (i.e., by the experimenter). On the other hand, if the particular levels to be used for an experimental variable are chosen randomly, then the levels of that variable are said to be "random." If the levels of all experimental variables are fixed, then the researcher is using a "fixed effects model." Conversely, if all of the levels of all of the experimental variables are random, then the researcher is using a "random effects model." Finally, if any of the variables differ (as to being fixed or random), then a "mixed effects model" is being used.

It is worth noting that the primary concern here is not with the actual number of levels of a variable that are chosen to be studied, but merely with how the particular levels used were selected by the researcher. This may, at first, appear as a foolish concern since one could easily maintain that the particular levels (especially if there are only two) arbitrarily "fixed" by the researcher might have occurred if the levels had been selected randomly. Nevertheless, the concern is real because the choice between fixed or random variables determines what inferences can be made (and how one tests the significance of one's findings). Hayes (1972) appropriately concludes, in his discussion of this problem, that "all the inferences made under (the fixed effects model) concern means (and differences among means)" while "the inferences made using (the random effects model) deal with ... the variance of the population of effects actually sampled by the experimenter."

As obviously useful as the results from studies using the random effects model might appear to be for the practitioners in the real world, there are those who have maintained that it is unneeded by psychologists. A. E. Edwards (1960), for example, stated, "There may be isolated instances in which (the random effects model) can be justified for a behavioral science experiment, but, in general, this model seems unrealistic. The fixed effects model and the mixed model seem to be much closer to the realities of experimental procedures in the behavioral sciences."

2.4 THE PRACTICAL LIMITATIONS OF ANOVA

Many writers perceive, in ANOVA, the possibilities for simultaneously handling many different factors (i.e., experimental variables). Bailey (1971), for example, stated "*The most important contribution of the analysis of variance is the notion of the multiple-factor experiments ... (which) permit the richness and complexity of behavior to emerge.*" It is, of course, true that ANOVA is theoretically capable of handling any number of variables, each of which might theoretically have a large number of levels. Such an experiment, if carried out, would, perhaps, permit the "richness and complexity of behavior to emerge." It is, however, as Bailey states, unfortunately, more of a "notion" than something that can be realized. In fact, it is but a promise, destined to remain forever unfulfilled. The most obvious reason for this conclusion is that as one adds more and more independent variables and more and more levels in each variable, ANOVA demands exponentially increasing amounts of data. A factorial experiment, for instance, with F factors and L levels in each factor requires L^F data cells. With small numbers of factors and levels, this does not present a serious problem. A three-factor experiment, for example, with four levels in each requires only 64 ($= 4^3$) data cells; not an insurmountable problem. But even an eight-factor problem with six levels in each requires over a million and a half data cells. A ten-factor problem with ten levels in each would require ten billion data cells! Even if a researcher could fill 500 data cells in an hour, it would take more than 2000 years to collect the data required. And this would be true only if the researcher could work 24 hours a day!

Of course, many clever designs have been devised for ANOVA which, somewhat, reduce the technique's insatiable appetite for data. But these designs cannot accomplish this without sacrificing some ability to do that which ANOVA originally set out to do (i.e., to test for the presence of all main effects and all possible interactions). The fact must be faced; ANOVA is a less and less effective tool as the research problem becomes more complex and interesting. ANOVA may always remain a "convenient" method for investigating a few levels of a few factors at a time, but Peters and Van Voorhis were correct in their original assessment

of ANOVA; constructive research for the complex, real world is just ready to begin where analysis of variance leaves off'

ANOVA was always predominantly concerned with "testing the null hypothesis," terminology now firmly implanted in our vocabularies for which we owe Fisher an immense debt of gratitude. This very proper concern over whether results obtained in a study could have occurred by chance alone may well be the chief contribution that usage of ANOVA has bestowed upon us. Cochran and Cox (1957) may have best stated why significance testing is important when they said, *"A useful property of a test of significance is that it exerts a sobering influence on the type of experimenter who jumps to conclusions on scanty data, and who might otherwise try to make everyone excited about some sensational treatment effect that can well be ascribed to ordinary variation in his experiment."* But, while the use of tests of significance is required by ANOVA, they are not the exclusive domain of that technique. Tests of significance are available and employed by virtually all statistical methods. Even if ANOVA had never been invented or were, now, to be totally abandoned, researchers would still have tests of significance available to them.

Others properly criticized ANOVA for not providing results in a form which more directly showed the extent of relationship between the various main effects or their interactions and the criterion variable. The familiar ANOVA source table emphasized and made readily apparent which main effects and interactions probably occurred by chance alone and which probably did not. The ANOVA table, however, as originally presented, did not offer a clear indication of what percent of the criterion variance was being accounted for by each main effect and interaction. Thus, the ANOVA table emphasized which, if any, conclusions could be drawn about the various effects, but did not directly show the extent of practical importance of those effects. In the past twenty years, progressively more ANOVA applications are actually being accomplished by computerized multiple regression/correlation techniques. Depending on the particular computer program used, the various "sums of squares" terms may now add up to one with the value of each indicating the proportion of total variance being accounted for by that particular effect. Tests of significance

provided by such programs will be identical to those which would have been obtained if traditional ANOVA computations had been carried out.

2.5 THE CENTRALITY OF THE PEARSON EQUATION

Realization that ANOVA is, in reality, merely a special case of multiple regression/correlation apparently eluded most psychologists for many years. Indeed, virtually all the important data analysis techniques can be shown to have their basis, in one way or another, in the Pearson product-moment correlation coefficient. This is no mere coincidence, however; it stems from the fact that the Pearson equation always finds a solution for one set of numbers to predict another set of numbers. The solution obtained always minimizes the sum of squared errors, regardless of the kind of numbers being used (e.g., ratio, interval, ranks, dichotomies, etc.). Wherry, Sr. (1984) provides a discussion of the Pearson equation for many types of data. Correlation coefficients, when squared, indicate the percent of variance that can be accounted for by using one variable to predict another. Multiple correlation is simply the use of more than one variable to predict a single other variable. Thus, multiple correlation can always be used to determine the amount of variance of a criterion variable which can be accounted for by various predictors. ANOVA certainly differs from multiple correlation in that it requires its main effect and interaction variables (i.e., it predicts) to be statistically independent from one another by the experimental designs it requires to be used.

It is vital for researchers to fully appreciate that the method of data analysis (regardless of the actual steps one goes through) in ANOVA obtains identical results to that of multiple correlation; both determine what portions of the criterion variance can be attributed to the various experimental (or predictor) variables. Thus, ANOVA does not provide a different way of analyzing data! Further, the significance tests used by the two techniques can also be shown to be identical. It is true, however, that multiple correlation (or multiple regression which is equivalent) is far more general than ANOVA even though both techniques use predictor variables to "explain" a criterion variable. ANOVA refers to the predictor set as the "experimental" or "independent" variables or

"factors," and refers to the criterion as the "dependent" variable. Multiple correlation is more general in that it does not insist that the variables in the predictor set be independent of one another, but it can be used when that is the case.

It is also worth pointing out that multiple correlation has a more than a fifty-year history of recognizing that each predictor variable may contain some error of measurement and that the weights derived for those predictors by multiple correlation is fitting not only the real covariation between the predictors and the criterion, but is also fitting the chance error in that sample's data. Because of this, the correlation, in a subsequent sample, between the actual criterion scores and the predicted criterion scores (based on prediction weights obtained for the first sample) typically will be somewhat lower than the multiple-R value obtained in the first sample. This phenomenon is called "shrinkage" in the multiple correlation literature. Wherry, Sr. (1931) worked out an equation for predicting this shrinkage even before ANOVA appeared on the scene. The shrinkage equation can be used to indicate that selecting (and weighting) all possible predictors can result in prediction equations which work less well for subsequent samples than prediction equations based on fewer predictors. For the psychologist concerned with the application of research to realistic problems, the issue of predicting results in subsequent samples is a practical matter which is central to the issue of being able to generalize one's results. These types of problems and their solutions have been almost totally neglected by users of ANOVA. Such concerns must, however, be brought into sharp focus when one desires to make inferences about the variance of the population of effects that can be expected in the real world.

2.6 THE ROLE OF EXPERIMENTATION IN RESEARCH

When one desires to investigate human behavior, two traditional approaches immediately come to mind: field studies and laboratory studies. Both have their advantages and disadvantages. In field studies, the researchers typically observe people performing real tasks under real conditions. Events upon which data are collected in field studies will, therefore, represent samples from the real subject populations and task

and environmental domains of interest to the researcher. However, it might take added time, effort, and expense to travel to the field sites. Cooperation from people performing the tasks of interest may not always be easy to obtain. It may be difficult to record performance data without unduly interfering with the actual phenomena of interest. Finally, it may be impossible to determine what events will actually occur during the period chosen for the study. Consequently, events which do occur during a given field study may not be those which were of the most interest to the investigator, and still worse, they may be unrepresentative of the entire domains of tasks, environments, and people of interest to which the researcher would like to generalize the results of the study.

Laboratory studies, on the other hand, offer the opportunity to control many, if not most, of the events which occur during the period of study. But these various events to be studied must be created and/or controlled. This, too, may be costly and time consuming. And, in a laboratory, it may be very difficult or even impossible to create certain situations which are sufficiently realistic to be perceived by the subjects like those in the real world. Nevertheless, the major advantages of laboratory studies reside in the ability to exercise control over the events which do occur. Woodworth and Schlosberg (1938, 1954) discussed three major advantages to being able to collect data under controlled conditions. These advantages are paraphrased below.

1. By controlling when events of interest occur, the experimenter can be fully prepared to observe and/or record the behavior being studied.

2. Because the experiment is controlled, the sequence of events which occurred can be known and can be repeated, if desired, by either the experimenter or others to validate the results obtained.

3. Because the experiment is controlled, experimental conditions can be systematically varied to determine concomitant variation in the criterion.

It is obvious that field studies will not have these benefits because, in such studies, the researcher is constrained to collecting data on the events which happen to occur. However, it is worth noting that each of these benefits of controlled experimentation can still be realized without insisting, like ANOVA does, that experimental variables be made statistically independent of one another. That is to say, ANOVA requires controlled experimentation, but controlled experimentation does not require ANOVA. This is a crucial point which is not immediately obvious to all researchers because many of them have closely associated controlled experimentation and the analysis of variance for such a long time.

Far too much confusion has been created in some textbooks on experimental design by various authors who shall remain nameless here. For example, some authors misdefine "independent variables" as "those which are controlled by the experimenter". By such a definition, all variables which are controlled by an experimenter must also be made independent of each other. This, in essence, requires that all experiments be of the ANOVA type. Instead, "controlled variables" should be defined as "those variables whose levels, during events in the experiment, are specifically determined by the experimental design process." Experimental variables, thus, can be either independent of or related to each other and still be controlled. The ANOVA experimental design process does require, of course, that all controlled variables be independent; the RSDV experimental design process, on the other hand, does not! Still other authors have gone so far as to make a grossly erroneous distinction between experimental data and correlational data, as if doing an experiment somehow precluded the obtaining of correlations or as if correlational data could not have been obtained from a truly scientific study. It is difficult to understand how the authors, themselves, can be so misguided, but it is even worse that their mistaken ideas appear in textbooks on experimental design and are being taught to unsuspecting students.

3. DECISIONS EFFECTING HUMAN PERFORMANCE STUDIES

When an investigator sets out to study various aspects of human performance, many decisions must be made. Making these decisions is not, in and of itself, a difficult task, but the decisions made can greatly influence the conclusions which are, can be, and/or should be drawn. Understanding the implications of one's decisions is, thus, of great importance. The decisions to be made fall into three general areas which shall be identified as:

1. elements of the situation to be studied,
2. measures of performance to be collected, and
3. methods of data analysis to be used.

The following sections give a brief introduction to the decisions which must be made in these three areas.

3.1 ELEMENTS OF THE SITUATION TO BE STUDIED

The most obvious decisions which must be made before a study of human performance can be undertaken concern deciding what situations will be studied. We have previously discussed some of the advantages and disadvantages of both field and laboratory studies. If one chooses to conduct field studies, then one is forced into studying the situations which happen to occur out in the field. In a laboratory, however, since all the situations must be created, investigators must decide specific elements of the situation that will be made to occur together so they may be studied.

Four major elements of situations are here identified. They are the specific:

- a. humans on whom data are to be collected,
- b. tasks with which the humans will be confronted,
- c. environments under which the humans will be performing, and
- d. times and sequences in which the humans perform the tasks in the environments.

The first three elements represent what shall be referred to as the **domains of populations** to be sampled by the investigator. The first three (e.g., humans, tasks, and environments) represent the generic domains found in all human performance studies. It must be obvious that no study can investigate all humans, all tasks, or all environments, so a given study must always be restricted to specific human, task, and environment domains. It seems fairly obvious that the researcher should be the one to specify which domains are the subject of a given study. The fourth element represents the intersections of the three specific domains during the various situations actually studied. Again, since it would be unlikely that all possible situations from the specified domains could be studied in a single experiment (unless the human, task, and environment domains are extremely restricted ones), the investigator must settle for some sample of the possible intersections. Thus, all experimental design decisions are actually concerned with either **domain specification** decisions or **domain sampling** decisions.

3.1.1 Definition of a "Domain"

A **domain** (or population) is a statistical concept which is, perhaps, best simply defined as including "all the possible instances of humans, tasks, or environments which are, were, or will be of interest to the researcher for a given study." The definitions of each domain can be as broad or as narrow as the particular investigator wishes to make them. There is, however, an impact which must be realized when the investigator selects and specifies the domains to be studied. The impact is that, as an honest and objective scientist, the researcher is only permitted to draw inferences and conclusions about the domains which are being studied.

Further, even this is permitted only when the samples drawn from the domains meet certain criteria.

It isn't that researchers cannot intelligently speculate about the probable similarities of their findings for domains other than those studied. Most of them are certainly capable of this. It is merely that any such claims they might make, at that point in time, are simply untested hypotheses, and, as scientists, they should never confuse the reader by mixing legitimate inferences that can be drawn from their data with untested hypotheses they think might be true. One may sympathize with the obvious desire many researchers have to generalize their findings and conclusions beyond what is actually warranted by their studies. They should remember, however, that use of appropriate procedures and statistics have already permitted them to make inferences about the domains actually studied based solely on the limited samples of data they have collected; conclusions about domains not studied are simply not justified. Still, it is doubtful that researchers can ever be convinced not to share with readers the benefits of insights gained during their studies. Nor should such behavior be discouraged too strongly. Insights and speculations may sometimes turn out to be valuable both to other researchers and to practitioners in the field. Unfortunately, much of what is often placed in the "findings and conclusions" sections go well beyond legitimate inferences, and, if included in research reports, should clearly be labeled as speculations.

3.1.2 Problems with ANOVA Studies

Earlier, it was indicated that all human performance studies must be studying a sample of some particular domains of people, tasks and environments. It should also be noted, however, that the particular domains studied by an experimenter may not necessarily have a sufficiently close correspondence to any actual domains of interest in the real world. In part, this has been why results from ANOVA studies have been unable to generalize to many real world situations. Regardless, having decided what domains will be studied, the researcher must next decide how those domains will be sampled. With complex domains, it will usually be impossible to study all possible combinations of all possible levels of all the

variables of all the domains. ANOVA properly recognizes the distinction between arbitrarily "fixing" the levels and "randomly" choosing them. The impact these decisions have on the type of inferences that can be legitimately drawn in those cases was previously discussed, but they bear repeating here. If the levels are arbitrarily selected, then no inferences are permitted about the domain in general. This is not simply a nice convention to follow; the tests of significance accomplished when the levels of variables have been arbitrarily selected do not reveal anything trustworthy about how much performance variance that variable may effect in the real world. This, then, is also a major reason that most "fixed" or "mixed" effects ANOVA design results from laboratory studies fail to generalize to the real world; they were not properly designed to reveal the practical importance of the experimental variables to the human performance measured. That so many studies over so many years turned out to be of little help in solving the problems of the real world should not surprise anyone. What is surprising is that so many researchers continue to behave as if the relative importance of variables, as found in their studies, will generalize to real world situations, when, in fact, there is no statistical evidence supporting that position. While it is true that researchers have been able to properly conclude that different levels of certain experimental variables probably do have a real effect on certain human performance criteria, the extent of effects found in laboratories may not be indicative of the extent of the effect in situations of interest in the real world.

Finally, it was previously pointed out that ANOVA unduly restricts both the number of experimental variables that can be investigated and the number of levels in each variable. For all practical purposes, it is impossible to effectively vary all the parameters which are needed to describe complex, domains of interest in the real world by using ANOVA. This type of restriction has forced the users of ANOVA to decide to investigate situations which are unlike those occurring in the real world of interest.

3.1.3 Further Difficulties With Field Studies

The field researcher does not have to make many of the decisions mentioned above. The people, tasks, and environments found in the field situations are obviously ones from the domains of interest. There is, in fact, much to commend researchers to go into the field and collect data there. But there are also several serious drawbacks, many of which have already been mentioned. In addition to those already mentioned, some others can be added.

First, when the study is completed, the researcher may not know if the particular sample of people, tasks, and environments were truly representative of the domains of interest. This is not a question about whether the events and behaviors observed were sampled from appropriate domains, but to what extent the obtained sample is representative of the distribution of situations for the people-task-environment domains as a whole. Assurance is needed that the tasks, which happened to occur while data were being collected, were neither too easy or too difficult. A field researcher needs to determine if the tasks, for example, are a representative cross section of the tasks which humans are required to do, not only at the field site(s), but at any other field sites of equivalent interest. Convincing evidence is also needed to demonstrate that the distribution of the levels of intelligence of the people on whom the data were collected is also a representative cross section of the persons' intellectual capacity found in the entire people-domain of interest. Similar reassurance is needed with respect to levels of experience and training and motivation. There are many such issues which should be resolved before the field researcher should claim that results obtained in a particular field study typify the results that would have been obtained if the entire domains of interest had been exhaustively studied. The major concern, then, is whether samples of data collected in the field situations are representative of the domains of interest, or whether they are, in fact, biased samples of data. Thus, the field researcher, like the user of ANOVA designs, may also have problems with generalizing results.

3.1.4 The Need for "Domain Specifications"

The prior discussions suggest a central, but often neglected, responsibility for both laboratory and field researchers; the development of specifications for the domains under investigation. Both types of researcher should embark on what can be called **domain specification** efforts, the purpose of which is to establish the relative frequencies with which various situations (i.e., combinations of types of people, tasks, and environments) occur in the entire, complex, real world of interest. Both types of researchers need domain specifications to explain, more precisely, to others what domains supposedly were investigated. The laboratory researcher should also have domain specifications prior to deciding how the domains of interest shall be sampled. The only way the field researcher can begin to determine if field data collected are representative of the "real" domains of interest would be to compare the actual sampled distributions of people, tasks, and environments with specifications of the distributions of people, tasks, and environments which define the domains of interest.

Conceivably, with such domain specifications in hand, field researchers could at least make estimates of the similarities between the sampled distributions and the specified distributions for those domains. If the means, variances, and interrelationships among the major variables are not significantly different from those same parameters for the specified domains, then the researcher would be justified in making generalized inferences to those domains. In other words, it is possible to develop statistical tests to determine if the situations occurring in a given field study differed significantly from those in the specified domains.

It should even be possible to develop procedures for weighting sampled data so that the overall results obtained from a particular field sample would be more representative of the specified domains of interest. The approach might have similarity to the equation for correction for curtailment of range. The concept of weighting data is somewhat foreign to many investigators. However, it should be remembered that "unweighted" data have all been assigned equal weights, and this, too, is an arbitrary

decision. Thus, there is really no escape from assigning weights; the only question is what are the most appropriate weights.

3.1.5 Inability to Resolve Unexpected Data Cases

A problem alluded to earlier concerns finding strange and unexpected performance values in data collected during a field study. The researcher often cannot tell if there is a problem with data recording devices or whether the data actually represent real behavior. Because, in field situations, it is usually difficult to monitor and record everything that is occurring, the researcher may not have recorded all the states of all the variables that could be effecting performance. Because of this, the researcher may not have a good idea of how to go about trying to replicate a particular situation to see if the same behavior will reoccur. Even if the field researcher returns immediately to the same site and obtains the same people to study, the same events may not occur and the researcher will be unable to resolve the issue. It is this lack of repeatability that is sometimes frustrating to those who venture into the field. Of course, similar experiences can also happen in the laboratory, but the degree of ambiguity is rarely as great in laboratory studies as in most field studies.

3.2 MEASURES OF PERFORMANCE TO BE COLLECTED

The second major decision area for an investigator concerns what measures of human performance are to be recorded. Traditional measures of performance include such parameters as speed and accuracy. If speed of performance is to be measured, the researcher must decide how to measure task duration and/or if times will also be recorded as various portions of the task are completed. A related problem is that it may not always be obvious when the task or a subtask has been completed. There is also the problem of what time to record if the task is not completed properly. Similar issues can also be raised for measures of accuracy of performance. Often, there may be difficulties in operationally defining what constitutes accuracy or being accurate, and there may be no equipment to objectively and directly measure it. The use of the experimenter or an outside "expert" to make these evaluations is fraught with problems. Making the human subject responsible for informing the researcher that he

has successfully completed the task may introduce yet new sources of variance that have little to do with actual performance. There is also the problem of whether partial credit should be given if the task is performed almost correctly. The temptation to dwell on these issues or to comment on various other types of measures which may or may not be closely related to task performance (e.g., physiological measures, subjective estimates of "workload," etc.) will be resisted. It is sufficient, for our purposes here, to point out that decisions regarding which performance data will be collected may, ultimately, lead to differing results, and, therefore, to different conclusions.

3.3 METHODS OF DATA ANALYSIS

The actual methods and procedures selected for the analysis of the data collected during a study can also impact conclusions drawn by an investigator. These data analysis methods and procedures include those for summarizing data, making inferences, and testing of hypotheses. It is helpful to discuss the more familiar methods and procedures in terms of their applicability to summarizing information about a single variable, two variables, or more than two variables at a time (i.e., univariate, bivariate, or multivariate methods and procedures). Statistics derived for these purposes typically provide single numerical values which describe, in a fairly unambiguous fashion, some important characteristics of the data which have been collected.

3.3.1 Univariate Measures and Methods

Univariate statistics help in describing the shape of the distribution of scores for a single variable. They can be computed and reported in lieu of publishing the raw data or providing a histogram showing the frequency of various scores. A single raw score (e.g., a person's time to perform a particular task under some specified conditions) would be fairly meaningless without some indication of how long it took other subjects to perform that same task under similar conditions. The mean, standard deviation, and the measures of skewness and kurtosis are four measures that provide highly meaningful, but different, summary information about the distribution of any set of scores

representing the obtained states of any experimental or criterion variable.

While the median and mode also provide interesting information about, say, how others did on a task, the mean has certain properties which make it particularly useful. For example, the mean minimizes both the sum of errors and the sum of the squared errors if we desire a single numerical value to "predict" what each score in the sample was. That is,

$$\bar{X} = \sum_{i=1}^N X_i / N, \text{ and} \quad (3-1)$$

$$\sum_{i=1}^N (X_i - \bar{X}) / N = 0, \text{ and} \quad (3-2)$$

$$\sum_{i=1}^N (X_i - \bar{X})^2 / N = \text{minimum.} \quad (3-3)$$

The mean is also useful because, if the sample of scores is a random one, then the obtained mean for the sample is also the "expected value" of the population from which that sample was drawn. What is meant by the expected value is that if we continued to draw samples of size N from the same population, the average of all the obtained sample means would equal the actual mean of the population. Thus, by computing the mean of a random sample of scores, we obtain an "unbiased" estimate of the mean of the population. This property of the mean allows us to draw inferences about the population mean, even though we have only looked at, perhaps, a relatively small number of cases from that population of scores.

The variance of a set of scores is computed by the equation

$$s_x^2 = \sum_{i=1}^N (X_i - \bar{X})^2 / N. \quad (3-4)$$

The variance provides a single measure of how the sample of scores vary around the mean of the sample. The square root of the variance is known as the standard deviation. The variance (and the standard deviation) also have very interesting properties when the sample of scores were randomly drawn from the population. If, for example, we were to continue to draw random samples of size N from that same population, we would eventually

find that the average of variances of the samples was slightly less than the variance of the population. In fact, we would find that

$$\text{Mean Sample Variance} = \text{Population Variance } (N-1)/N . \quad (3-5)$$

Thus, an unbiased estimate of the variance of the population can be obtained simply by multiplying the sample variance by $N/(N-1)$. This is true, regardless of the shape of the distribution of scores in the population. It can also be shown that the variance of the means of the samples will be equal to the population variance divided by N . Equations (3-4) and (3-5) are part of the basis for testing whether two or more random samples probably came from the same population, and are, thus, part of the basis for ANOVA. What is particularly important, however, is that Eqs. (3-4) and (3-5) are based solely on random sampling and not on the shapes of the distribution of scores in the population from which the random samples are drawn. The assumptions in ANOVA (and many other procedures requiring statistical tests) which concern normal-shaped (the so-called bell-shaped) distributions are there in order to justify using the F -table values (which are based on random sampling from normal distributions) to determine how often certain results would have happened by chance alone. But, equations (3-4) and (3-5) still allow researchers to draw inferences about the variance of the population and the variance of the means of samples from that population, regardless of the shape of the distributions from which a data sample was drawn, provided the sample was drawn randomly.

Measures of skewness and kurtosis are found, respectively, by obtaining the ratio of the average of $(X_i - \bar{X})$ raised to the third and fourth power and the standard deviation raised to a similar power. The skewness measure provides information about the sample of scores in terms of whether the sample distribution tends to be symmetric around the mean while the measure of kurtosis provides information about the flatness or peakedness of the distribution of a sample's scores. Of particular interest, here, is that all symmetric distributions will have a skewness value of zero. A negative skewness value would indicate a longer "tail" of lower-valued scores while a positive skewness value would indicate the

opposite. With regard to the measure of kurtosis, a normal distribution will have an expected value of 3, while a rectangular distribution (i.e., one where there is equal probability of randomly drawing a score along the entire distribution of possible scores) will have an expected kurtosis value of 1.8.

While measures of skewness and kurtosis are less frequently reported on sample data, they do provide very useful information about the shapes of the distribution of the sample scores.

3.3.2 Bivariate Measures and Methods

Bivariate procedures and statistics provide information which allows the researcher to reveal some properties about two different variables, or more precisely, two different sets of scores which supposedly measure different properties of the elements in the sets. The Pearson correlation coefficient (r), which can be correctly interpreted in many different ways, is computed by the equation

$$r_{x_1 x_2} = \frac{\sum_{i=1}^N z_{1i} z_{2i}}{N} \quad (3-6)$$

where

$$z_{1i} = (X_{1i} - \bar{X}_1) / \sigma_{x_1},$$

$$z_{2i} = (X_{2i} - \bar{X}_2) / \sigma_{x_2}, \text{ and}$$

N = the number of cases in the sample.

One interpretation of r is that it gives the slope of the prediction line that minimizes the sum of the squared errors when both variables' sets of numbers have been converted to standard scores. That is, the "best" linear prediction of either set of standard scores will be found by multiplying the other set's standard scores by r . It, therefore, gives a summary of how two variables' scores vary together. The square of its value also yields the proportion of variance of either of those variables that can be accounted for by using the other's set of numbers to predict it. If the sample of N cases was obtained by random sampling, then the

obtained value for r is also an unbiased estimate of the population's correlation for those two variables. Again, it needs to be emphasized that this conclusion is not dependent on any assumptions about the shape of the population distributions, but simply a property of random sampling.

With regard to the correlation between two variables, tests of significance can be performed to give an indication (provided the assumptions of the tests are adequately met) of whether the correlation differs significantly from some particular value (e.g., zero or some other value) by an amount greater than might be expected by chance alone. Thus, the simple correlation coefficient can be used to test the null hypothesis that some variable (which describes some feature of one of the domains of interest) had no effect on a variable which is some measure of human performance. While the various tests of significance for a correlation coefficient typically do assume some particular shaped population distribution (e.g., normal, binomial, or rectangular), it can be shown that these tests tend to converge toward the tabled values (F-table) for normal distributions when N approaches 30 or more cases.

A rather large number of methods have been developed for comparing two data samples to decide whether they probably did, or did not, come from the same population. Examples of these methods include various standard t -tests, Mann-Whitney U-test, White R-test, Festinger d -test, etc. In reality, these methods are simply special variations of determining if two sets of numbers (where scores in one of the sets are measuring some variable and scores in the other set are either "one" or "zero" to indicate if they came from sample one or not) can be said to be significantly related. Thus, these methods actually fall into the category of correlational methods.

3.3.3 Multivariate Methods: Multiple Correlation

Multiple correlation (perhaps the most widely used of the multivariate procedures) represents a sequence of tests of the null hypothesis that begins by assuming the null hypothesis with regard to all the potential predictors (i.e., the variables which describe features of the domains of interest). The predictor variable having the largest

correlation with the criterion (i.e., the measure of performance) is then tested to see if the null hypothesis can be rejected. If so, then its effect on the criterion is said to be "real" and that predictor's effect on the criterion and all other remaining predictors is removed (i.e., "partialled" out). The above procedure is repeated for all the remaining potential predictors until the largest remaining residual correlation between the criterion and any of the remaining predictors is no longer sufficiently large to reject the null hypothesis by the particular test of significance employed. This approach to multiple correlation is called the "accrual" method (because it keeps adding predictors one at a time). Wherry, Sr. (1940) developed the first of the accrual methods for multiple correlation using his "shrinkage" equation to determine when to stop selecting predictors so as not to overfit the errors of measurement in the particular sample of data that had been collected.

Another approach to multiple correlation is called the "deletion" method. It starts by selecting all available predictors and then deletes the most nonsignificant one by the particular test of significance being employed. This procedure is repeated until none of the remaining predictors are able to be rejected as being nonsignificant. The "accrual" method and the "deletion" methods can sometimes arrive at a slightly different set of "significant" predictors, but, usually, they obtain identical results when both use the same test of what is deemed to be significant. The accrual method requires far fewer calculations, however. Either method will also usually obtain a slightly different solution (i.e., fewer or more selected predictors) depending on a level (e.g., .05, .01, etc.) arbitrarily selected by the researcher. Since all research results might have occurred by chance alone, some level of probability must be chosen to define a point beyond which the researcher is willing to identify the results as probably not being merely the result of chance.

The third approach to multiple correlation is to simply use all the available predictors, but this, because of the shrinkage problem, will typically lead to less satisfactory solutions since the weights obtained for the predictor variables will not work as well in a cross-validation

sample. Finally, it should be pointed out that the sequence in which variables are selected will make no difference in either the final value of the multiple correlation (R) or in the weights assigned to each variable selected as a predictor. What is important is the set of predictors selected; not their sequence.

While multiple correlation is always finding the "best" linear fit between selected predictor variables and a criterion, the researcher is permitted to define other variables which are nonlinear functions of a predictor variable or the products of various predictor variables. Further, nonlinear transformations can be accomplished for any criterion variable prior to the accomplishment of the multiple correlation procedure. Thus, multiple correlation can also be used to identify significant nonlinear and interaction effects as well as linear ones. When such variables are used, it is traditional practice to select all possible significant linear variables before attempting to select the nonlinear or interaction predictor variables. Stone and Hollenbeck (1984) have recently discussed issues surrounding the question of the sequence with which predictors should be selected. Thus, when using multiple correlation to accomplish ANOVA applications, the main effect variables are selected first, then interaction terms. ANOVA, however, essentially uses the "deletion" method discussed earlier in that it, first, selects everything and then determines which predictors are probably not significant ones. ANOVA may "pool" the variances of nonsignificant effects with the error (i.e., the "unexplained" variance) term.

While there is general agreement with regard to the sequence in which main effects and interaction effects should be selected, the sequence of selecting various linear and nonlinear terms for the main effects themselves represents another potential point of disagreement among researchers. If a main effect has k levels, then ANOVA (using multiple correlation to accomplish the analysis) requires $k-1$ predictor variables to represent the total main effect variance. The reason this is true is the same reason that the main effect has $k-1$ degrees of freedom. Traditional ANOVA will select all of these $k-1$ predictor variables, first, for the major analysis, and only, subsequently, may determine by "tests of

contrasts" whether there are significant differences among those $k-1$ variables. In this approach, the $k-1$ variables are dichotomous variables (zero or one scores) indicating that the data case belonged or did not belong to that level of that main effect. Alternative procedures in ANOVA for main effects in which the particular levels can be assigned numerical values, utilize $k-1$ orthogonal polynomials to determine the probable significance of linear, quadratic, cubic, and higher-order terms. As in traditional multiple correlation, it would seem more appropriate to, first, determine the significance of all linear terms of all main effects prior to determining if any quadratic or higher-order polynomials are significant or not.

3.3.4 Multivariate Procedures: Factor Analysis

Factor analysis is another widely used multivariate procedure which has many variations. Its primary function is to determine the number of independent dimensions (or factors) necessary to account for the interrelationships among a set of variables. Several methods have been developed to determine how many independent factors are needed or should be extracted. Various schemes have been devised to rotate the dimensions obtained to various mathematical criteria in order to obtain a "meaningful" set of independent factors. The most popular rotation method among psychologists is the Varimax method which attempts to rotate to mathematical criteria which will yield a pattern of factor loadings which exhibits what is referred to as "simple structure." A recent improvement to that method which attempts to find a pattern of loadings which exhibits not only "simple structure" but also "positive manifold" is discussed in Volume 2 of this series. Both "simple structure" and "positive manifold" are concepts which help to determine the rotation which represents the "real" factors. A different "rotation" represents an alternative method for accounting for exactly the same variance, but with the "factors" being located in somewhat different positions. The relationships of the variables in the matrix of interrelationships to the obtained (and/or rotated) factors which were found are called "loadings." Because the factor loading (f_{ik}) is, in fact, a measure of relationship between a given variable (i) and a given factor (k), the loading squared will indicate the proportion of variance of that variable which is accounted

for (or "explained") by that factor. A performance criterion variable may be found to load on several different independent factors. Interpretation of what a particular factor represents is normally based on how each of the variables in the matrix actually load on that given factor.

From the above discussions, it can be seen that multiple correlation can be used to test hypotheses regarding significant relationships between the domain descriptor variables and the human performance (criterion) variable. It can also be used to determine how much of the criterion variance can be "explained" or "predicted" by the predictors. The analysis of variance (ANOVA) can now be recognized as a special case of multiple correlation which can be used when, and only when, the predictor variables have been mathematically forced to be independent of each other (regardless of how they may be related in the real world). Factor analysis, on the other hand, can be used to determine how many significant independent dimension or factors are responsible for the interrelationships of both domain descriptor variables and/or performance criterion variables, and to further determine the nature of the factors and the extent to which each factor accounts for the variance in performance scores.

4. PRINCIPLES AND APPLICATIONS OF RANDOM SAMPLING

4.1 RANDOM SAMPLING DEFINED

Random sampling does not, of course, refer to "careless" or "haphazard" sampling. Random sampling concerns the probability each element of a "population" has of being selected on the next draw from that population. If all the population's elements truly have an equal chance of being selected, then, and only then, is the sampling said to be a random one; any departure from equal probability for all elements in the population is some form of "biased" sampling. The entire sample drawn may be said to be a "random sample" provided each draw has been random. In a random sample, therefore, each draw is independent of what elements have previously been drawn from the population. These features (i.e., equal probability of being drawn and independence of each draw) permit calculations of the "expected" values of various statistics used to describe samples.

4.2 "RANDOMIZATION" IN EXPERIMENTS

In section 2 of this report, R. A. Fisher's ANOVA was criticized as being, in part, responsible for slowing psychology from becoming a viable scientific discipline capable of coming to grips with the complex problems of the real world. To be sure, the adoption and widespread use of ANOVA has made, and will continue to make, vital contributions to the various fields of psychology. Nevertheless, its utility is questionable for studies of human performance in which the researcher suspects that a large number of variables are responsible for the variation in human performance. While ANOVA obviously falls into the class of multivariate procedures (because it can be used to investigate the effects of more than one variable at a time), researchers should recognize that ANOVA has some very serious, practical limitations. Fisher must have realized that ANOVA,

with its demand for proportional data cells for its main effects, could not be used to systematically vary a large number of experimental variables at the same time. As a scientist, he recognized that during any experiment, some "uncontrolled" (and, perhaps, even unknown) variables could be responsible for some of the variation found in the criterion or dependent variable. The fact that all possible variables which might effect the criterion could not be controlled by ANOVA was probably not a major concern to Fisher. Fisher, no doubt, wrestled with the problem of how to alleviate the possible covarying of controlled and uncontrolled variables. And out of that concern came one of his most significant contributions to the experimental method; the use of the device called "randomization."

Fisher was a brilliant statistician, steeped in the concepts of random sampling. He thoroughly understood that when two truly independent variables were randomly sampled the expected value of the covariance would be zero. Thus, he correctly reasoned that the best way to limit the covariation of uncontrolled and controlled variables would be to randomly assign persons to the various experimental conditions. Randomization, in experimentation, is often thought of as a needed "insurance" policy to prevent the researcher's subconscious biases from intervening into his experiments, or as a way of keeping other possible biases, which otherwise would unduly influence the results of the experiment, from occurring. In a loose sense, randomization serves these functions in that, by randomly assigning people to groups, treatments, etc., it tends to reduce the probability of certain unusual combinations of persons, tasks, and environments that would yield results that others, performing essentially the same experiment, would not be able to replicate and substantiate.

The importance of randomization to experimentation cannot be overemphasized. Cochran and Cox (1957) have stated, "*Randomization is one of the few characteristics of modern experimental design that appears to*

be really modern. One can find experiments made 100 or 150 years ago that embody the principles that are now regarded as sound, with the conspicuous exception of randomization."

The concept of randomization is but one application of the principle of random sampling which is the basis for almost all statistical tests. The principle of random sampling also permits establishing confidence limits, the derivation of "unbiased" estimates of various statistical parameters, the testing of hypotheses about samples of data, and the drawing of inferences about the populations from which samples of data must have come. The principle of random sampling is absolutely vital to statistics and, thus, to experimental design and the analysis and interpretation of data.

4.3 STATISTICS BASED ON RANDOM SAMPLES FROM A POPULATION

While it is true that many of the tests of significance are derived by assuming random sampling from normally distributed variates, it must be pointed out that the principles of random sampling apply to any type of distribution. The various statistics which measure the mean, standard deviation, skewness, and kurtosis can be calculated for any type of scores and from any shaped distribution. Regardless of the shape of the distribution of a population of scores, the mean of a random sample from that population can be shown to provide an unbiased estimate of the mean of the population. Likewise, it can easily be demonstrated that an unbiased estimate (or "expected" value) of the variance of a population of scores will be the sample's variance multiplied by $n/(n-1)$, provided, again, that the sample of n scores were drawn at random from the population. In the same way, the correlation of randomly sampled scores between any two variables (of whatever shaped distributions) can be shown to be an unbiased estimate of the correlation of those variables in the population. These properties of various statistics based on random sampling allow researchers to make valid inferences about the population from which only a single sample of data has been drawn.

4.4 THE STATISTICS OF MULTIPLE RANDOM SAMPLES

One of the most intriguing of all statistics deals with the expected variation of the means of different samples of size n , each of which has been randomly drawn from the same population. It can be shown that the expected variance of the means must be equal to the variance of the population divided by the size of the samples. That is

$$\sigma_{\text{MEANS}}^2 = \sigma_{\text{POPULATION}}^2 / N. \quad (4.1)$$

It is this property of random samples which permits the testing of the significance of the difference(s) between the means of two (or more) samples. While it is true that the values found in various tables (e.g., F , t , Chi Square, and z) are all based on random sampling from normally distributed populations, equation (4.1) holds for any shaped distribution.

4.5 RANDOM SAMPLING OF ITEMS WITH MULTIPLE ATTRIBUTES

When random sampling is discussed in mathematical or statistical textbook examples, the population of items to be sampled are usually described as varying on only a single attribute. Thus, most textbook examples describe a population of different colored balls or some other single variate. Typically, textbook examples also assume some well known theoretical distribution shape such as normal, rectangular, or binomial. However, we may equally well consider random sampling from a population of items that vary, not only in color, but also in size and shape as well. Indeed, we may think of each item in the theoretical population as varying simultaneously on a multitude of different attributes. Further, we may also consider that each of these attributes may possess different distribution shapes, none of which are perfectly normal or rectangular. If those items (with multiple attributes) are randomly sampled, we certainly could ignore all the states of all other attributes except for a particular attribute of interest. It should be obvious that, assuming the sample drawn is a truly random one, then we would be able to calculate unbiased estimates of that particular attribute's population mean, variance, and so forth. But if that is true for the particular attribute we happened to be interested in, then it would also be equally true for all other attributes as well. Thus, the sample obtained from a true

random sampling of multi-attribute items of a given population or domain will yield a random sample for each of those attributes. The fact that some of the attributes of those items may not be independent of one another does not detract from the above conclusion. And, the conclusion will also be true whether the random sampling is accomplished with or without replacement of each item drawn. That is to say, for the statistics of random sampling to hold, the attributes of items are not required to be independent of each other, but the drawing of each subsequent item must be independent of all previous draws. Random sampling is totally satisfied when each item remaining in the population has an equal chance of being selected on the next draw. If, for example, attribute A and attribute B are related in the population to be sampled, the correlation obtained from a random sample of those items should also provide an unbiased estimate of the correlation of those attributes in the population. Similarly, the obtained means for both attribute A and attribute B should both be unbiased estimates of the means of those attributes in the population.

When experimental variables are spoken of as having effects on some performance criterion, one is merely stating that there are significant relationships between those experimental variables and the criterion variable. These relationships are assumed to be stable in the population. Because researchers do not know the extent of those relationships in the population (or domains of interest), they must collect data so as to obtain estimates of those relationships. As pointed out earlier, the only way to ensure that the estimates obtained are unbiased ones is to ensure that the sample of data collected is truly a random sample of such data. This philosophy is the basis for the Random Sampling of Domain Variance (RSDV) technique.

4.6 RANDOM SAMPLING IN SIMULATION AND MODELS

Another practical application of random sampling is found in various simulations and mathematical models. Often times, investigators desire to mathematically replicate conditions found (or assumed to exist) in the real world so as to test theories or to artificially create situations in laboratories so as to be able to present realistic

situations to persons (for training or other purposes). Many times, investigators will be aware that, in some domain of interest, some particular conditions occur with a given frequency. For example, it may be known that some condition (A) occurs eighty percent of the time and the alternative condition (B) occurs the other twenty percent of the time. Because of incomplete understanding of what causes those conditions to occur or not, the investigator does not know when to use condition A and when to use condition B. To overcome this problem, the simulation is designed to generate a random number between zero and one to help make this decision. If the value of the generated random number is .8 or below, condition A is used, whereas condition B is used if the generated random number is greater than .8.

This approach of using random numbers (usually generated by a computer) to decide among two or more alternative conditions which have differential probabilities of occurrence has been widely and successfully used for many years in a large variety of simulations. The RSDV concept incorporates this approach in experimental studies for determining what tasks and environmental variable states will be used for collecting performance data on a particular subject during a particular trial.

4.7 THE WHERRY, JR., SIMULATED DATA GENERATION TECHNIQUE

Another application of random sampling combined with simulation is a technique developed in 1962 by this author for generating samples of fictitious, or simulated, multivariate data samples each of which possess characteristics (i.e., means, standard deviations, and correlations) similar to the characteristics of known real data. To a large extent, this particular technique is the most immediate historical antecedent to the RSDV technique. It may be helpful to the reader to understand the particular research problem out of which the simulated data generation technique first emerged. For this reason, the research problem is briefly discussed in the following paragraphs; a more complete discussion of the problem and its solution can be found in Wherry, Sr., Naylor, Wherry, Jr. and Fallis (1964).

4.7.1 Generating Simulated Rating Data

The basic research question was concerned with determining if different strategies were being used by different military raters in deciding upon their overall evaluations of their subordinates. More precisely stated, the problem involved a determination of the extent to which various performance aspects were felt to be differentially important by the raters. Typically, military personnel are routinely rated on a number of standard variables. These ratings become part of the permanent records for those individuals. A person's ratings may later be used to help determine which of those ratees is to be promoted, selected for a particular assignment, nominated to be sent to a particular school, and so forth. Obviously, some rated variables will be more important than others for these different purposes. But even for the same purpose, various raters or rating reviewers may differ as to how important the different rated variables should be. To the extent that this is true, the raters or reviewers can be said to be employing different strategies in their overall evaluations, even if there were complete agreement among the evaluators as to how the ratees should be scored on each of the separate variables. Investigation of this particular problem was further complicated by the fact that scores assigned on the rated variables were known to be related to each other to varying degrees. Here, then, was a perfect example of the important "experimental" variables which undoubtedly influence the performance of the evaluators being related in the real world. It would have been foolish and unrealistic to pretend these variables were unrelated simply to be able to design an ANOVA study to investigate their effects on the performance of evaluators.

It was reasoned, however, that if a large number of raters were asked to make an overall evaluation for each member of a fairly large group of ratees (based on preassigned realistic ratings), sufficient data would then be available for determining if different strategies did, in fact, exist, and for comparing the relative importance of the different variables for each rater. In actual practice, none of the raters had supervised the same subordinates. Consequently, no data existed on which the different raters could be directly compared. What was needed, then, were realistic sets of ratings on a fairly large sample of ratees. It

should also be obvious that it would also be desirable for that sample of ratees to be a random one from the overall domain of ratees. Such ratings could be presented to the evaluators in question and, using them, they could be asked to give their overall evaluation of each ratee. While no such actual data existed, historical data were available on the means and standard deviations of the various rated variables and on how those ratings correlated with each other.

4.7.2 The Method for Generating the Simulated Data

To generate a random sample of ratings, the established means, standard deviations and interrelationships were used as specifications for the domain of interest. The matrix of interrelationships was factor analyzed to establish both the rating variables' loadings on each of the independent common factors needed to explain the interrelationships and each variable's communality (i.e., the percentage of the variance of each rated variable that was being explained by the common factors which had been found). Next, additional independent factors were created to account for the "unique" portion of the unexplained variance of each rating variable. This can be accomplished in two ways. Either a single "unique" factor is created for each variable (where its only nonzero loading is for the variable in question and is equal to the square root of the quantity one minus that variable's communality), or both a "specific" factor and an "error" factor is created for each variable. The latter procedure would be used when the "reliability" of the simulated scores is an issue in the study. In either case, the sums of the squares of the independent factor loadings across any variable will now sum to one and the sum of the products of the respective values in any two variables will yield the correlation between those two variables. To simulate a single ratee's ratings, a vector of normal random deviates is then generated by a computer. This vector represents that individual's standard scores on each of the independent factors. The standard score for any rating variable for that individual is then obtained by summing the product of that individual's normal random deviates multiplied by their respective loading on the independent factors. To obtain the raw score rating for that simulated individual on that given variable, the simulated standard score is multiplied by that variable's known standard deviation and then

that variable's known mean is added to it. In this way, fictitious scores could be generated for any number of simulated ratees to be used as the stimulus material for the evaluators. Of particular interest is that, as the number of fictitious cases generated using this method gets larger and larger, the means and standard deviations of the fictitious variable scores should get ever closer to the true means and standard deviations of those variables, and the correlations among the variables will get closer and closer to the actual correlations of those same variables. Regardless of how small or large any sample of cases (generated in this way) is, each sample is actually an unbiased sample of cases from the specified population.

Because of the speed with which computers can perform these operations, hundreds of samples of simulated persons can be created in very short times. Despite the obvious utility of this approach, a basic restriction of it was that it assumed normally shaped distributions underlying each of the rating variables. In the original application of the technique, this assumption was felt to be warranted; nevertheless, the technique was still far more restrictive than it needed to be. The RSDV technique incorporates many of the good features of this technique, but makes no requirement on the shapes of any of the underlying distributions of levels for any task of environmental variable of interest.

4.7.3 Other Applications of The Simulated Data Generation Technique

The fact that the original simulated data generation method did assume underlying normal variates made it particularly applicable for use in a variety of studies to test the efficacy of different statistical approaches and tests which, themselves, already include assumptions of normality of data distributions. These applications were usually in the realm of testing and selection problems. For example, Hutchins (1970) used this technique to investigate the efficacy of a multiple-battery approach for test-selection applications when the number of predictors is relatively large and the sample size is relatively small. Lane (1971) also used the technique to compare the Wherry, Sr. shrinkage equation with alternative ones proposed by Nicholson (1960) and Darlington (1968).

The technique has also been widely used in generating large samples of simulated persons whose anthropometric measures conform to actual, established anthropometric measures of known populations (e.g., U.S. Navy pilots, U.S. Air Force pilots, etc.). These simulated persons are then used to determine the percentages of such populations which will be unable to reach various controls in a workstation (e.g., a pilot's cockpit in a particular aircraft). The simulated data generation technique, thus, has also been useful in the solution of human engineering problems as well as in solving problems in the testing and selection realms.

In these subsequent applications of the simulated data generation technique, the simulated data were never presented to subjects in a study (e.g., the raters in the original application of the method), but were, instead, manipulated by computers for various purposes. The technique was so obviously applicable for rapidly generating unbiased samples of fictitious multiple variable data which possessed normally shaped distributions, that it was primarily recognized as a technique for simulating samples of people. It immediately became very appealing for the investigation of the efficacy of various statistical methods and procedures. Because of the obvious acceptability to those realms, and because of its dependence on normally shaped distributions, its original purpose of generating the actual stimuli to be presented to subjects in an experimental study was never considered to be one of its strong points.

The RSDV technique, however, completely removes all restrictions on the underlying shapes of the variables' distributions. Because of this, the variables being simulated can equally well be any task and/or environmental variables and need not be restricted to variables which describe human's capabilities (which, in general, do have normally shaped distributions). The RSDV concept can, thus, be thought of in terms of a generalized experimental design procedure for deciding what stimulus conditions will be presented to subjects on various trials of any experiment. Its basic strength and power derives from its reliance on random sampling theory. Because of this, by the end of the data collection period the experimenter will have randomly sampled the specified domains

of interest and, thus, will be able to use those data to compute unbiased estimates of the performance means and variances for those domains as a whole, regardless of how complex those domains may be. The procedures for conducting RSDV studies are discussed in detail in the following section.

5. THE RANDOM SAMPLING OF DOMAIN VARIANCE (RSDV) TECHNIQUE

5.1 THE OBJECTIVES OF THE RSDV TECHNIQUE

The Random Sampling of Domain Variance (RSDV) technique has several main objectives. First, it has as one major objective the providing of a methodology for the simultaneous investigation of multiple variables, each of which may have many different levels, in a controlled experiment. Unlike ANOVA, which has severe limitations on the total number of variables and levels within variables that can be investigated in a single study, the RSDV technique is, essentially, unrestricted. The second, and equally important, objective of the RSDV technique is to allow the results obtained from RSDV studies conducted in a laboratory setting to generalize to the real world. A third objective of the RSDV technique is to serve as a theoretical bridge for moving between laboratory and field studies.

In an earlier section, the advantages of controlled experimentation were enumerated. The RSDV technique permits all of those advantages to be fully realized in that: (1) the events of interest which will occur can be controlled so that the experimenter can be fully prepared to observe and/or record the behavior being studied; (2) the sequence of events which occurred can be known and can be repeated, if desired, by either the experimenter or others to validate the results obtained; and (3) the experimental conditions can be systematically varied to determine concomitant variation in the criterion (or criteria). Later, it will be seen that the method used for systematically varying the experimental variables in RSDV studies is quite different from the method normally used in systematically varying the main effects in an ANOVA design.

5.2 THE EFFICACY OF RANDOM SAMPLING

A basic difference between RSDV and ANOVA is that ANOVA typically uses repeated and exhaustive sampling of the same few levels and few variables while RSDV utilizes a far more efficient random sampling of any number of levels for any number of experimental variables. That is, ANOVA tends to require the experimenter to repeatedly sample the same levels of variable A for every level of variable B used in the experiment. RSDV does not require this for two very good reasons. First, and, perhaps, most importantly, the actual variables A and B may not be independent of each other in the real world. If they are not independent in the real world, then the ANOVA strategy forces the experimenter to collect data which are, at best, not particularly representative of situations in the real world, and, at worst, may even be drastically misrepresentative of the real world. If the ANOVA method forces the collecting of data which are not representative of the real world, then estimates of performance based on those unrealistic situations cannot lead to unbiased estimates of performance in the real world. Secondly, by randomly sampling more of the possible (and probable) combinations of A and B, the experimenter ensures an unbiased estimate of both of those variables and any concomitant performance variation they may be responsible for in the criterion. When the actual population of effects of interest consists of multiple variables and multiple levels within those variables, exhaustive sampling of all possible combinations may be neither prudent nor feasible whereas a random sampling of those possible combinations (in proportion to their likelihood of occurrence in the real world) will always be possible, obviously more efficient, and certainly more prudent since such samples do lead to unbiased estimates of the performance of interest.

5.3 PROCEDURES FOR CONDUCTING RSDV STUDIES

The capability of the RSDV technique to provide a means by which answers to complex and interesting real world problems can be obtained from experimentally controlled laboratory studies dictates a need for a systematic method for accomplishing such studies. There are four major phases involved in conducting RSDV studies. They are:

1. specifying the domains of interest
2. selecting the sample of situations to study
3. creating the situations and collecting the data, and
4. analyzing the data and drawing inferences and conclusions.

The third phase (i.e., creating laboratory situations and collecting data) is no different for RSDV studies than for other types (e.g., ANOVA) of studies. Consequently, no further discussion of that phase will be provided here. Also, the fourth phase (i.e., analyzing the data and drawing inferences and conclusions) presents no novel problems for RSDV studies. An RSDV study, of course, eliminates the possible use of ANOVA as the analytical approach and requires usage of one or more of the classical multivariate techniques (e.g., multiple correlation, factor analysis, etc.). Some possible approaches to how these multivariate techniques could be used to analyze the data from RSDV studies were previously discussed.

The first two phases of RSDV studies represent new kinds of activities not required by ANOVA studies. To a large extent, the activities required in these phases are very closely related to modeling and simulation. Many researchers may be unfamiliar with this field. The following sections provide some suggested approaches for how one can efficiently and effectively meet the requirements for accomplishing the activities required during the first two phases.

5.3.1 Specifying the Domains of Interest

Many possible domains of interest exist in the real world, and a researcher is free to choose the particular combinations of people, tasks, and environments to be studied. A given researcher's concern with some real world problem should dictate what particular categories of people, tasks, and environments are to be specified, but the scope of these domain categories may range from being quite narrow to very broad. The requirement to specify the particular real world of interest by specifying the people, task, and environmental domains is needed to resolve possible ambiguities of what is actually being studied. It must be recognized, however, that it is probably impossible to ever completely specify all

possible details about a given real world of interest. It is highly likely that no researcher really knows or understands all the details about some real world of interest. For this reason, a distinction can be made between the actual real world (ARW) of interest and the specified real world (SRW) of interest. A researcher can discuss the ARW in very general terms. For example, a researcher might typically say that he is interested in studying "the performance of assembly line workers who have to make rapid decisions under noisy conditions" or "the performance of military tactical officers who must assess and evaluate complex combat situations during periods when extreme danger is imminent". We may note that both of these statements contain information about the people-domain of interest (i.e., "assembly line workers" and "military tactical officers"). Both contain information about the task-domain of interest (i.e., "make rapid decisions" and "assess and evaluate complex combat situations"), and both contain information about the environmental-domain of interest (i.e., "under noisy conditions" and "during periods when extreme danger is imminent"). But, in general, while such statements may provide a preliminary global definition of the ARW, they tell others precious little about the specific composition of the respective people, task, and environmental domains in those ARWs. More to the point, however, is the fact that merely studying, say, some assembly line workers making some rapid decisions under some noisy conditions will probably not be a random (or representative) sample of that total domain. If that is true (and there is no reason to believe that it is not), then the investigator should not try to generalize his results to the larger, more inclusive domains which were only globally stated.

To a large extent, this problem is identical to the one concerning random and fixed models in ANOVA. The RSDV procedure does not demand that the investigator specify the people, task, and environmental domains in which he is interested. But, the RSDV procedure does insist that the investigator should fairly precisely specify the people, task, and environmental domains that were randomly sampled during the course of the study. Obviously, it is to those domains, and only those, which the obtained results can be generalized with any great degree of confidence.

These specifications begin with the names of the variables which describe the composition of the people, task, and environmental domains that will be sampled in the study. Such variables will hereafter be referred to as the **domain descriptor variables**. Thus, in reporting an RSDV study, the researcher should include a section entitled "Domain Descriptor Variables." Three subsections should be included for the people domain, the task domain, and the environmental domain. The nature of domain descriptor variables should be such that every element within the particular applicable domain must be able to receive a score on each named variable. For example, each person in the people domain must be able to be scored on each people domain descriptor variable. Similarly, every type of task which is considered to be part of the task domain of interest must be able to be scored on each task domain descriptor variable, and so forth.

Having decided on the various variables to be used as domain descriptors, the next step is to determine or estimate the probable distributions of the possible scores on each domain descriptor variable in that portion of the real world in which the researcher is interested. For example, with regard to the people domain, personnel records containing many different variables may already exist on actual persons in the people domain of interest. Prior analyses of such data may already be available which provide information on the distributions of those variables' scores and on the correlations among those variables. If not, a random sample of data from those personnel records could be obtained and such information could be calculated. If no prior data has been collected on the people in the people domain of interest, it certainly indicates that the researcher probably knows very little about the elements of that domain and its overall composition. If such is the case, it is certainly advisable for the researcher to take time to decide what variables would adequately describe the domain of interest and then to gather some actual domain descriptor variable data on a random sample of people from that domain of interest before going on with the study.

For the task domain of interest, it is even more likely that no records will exist which can be used to locate various tasks of interest

on task descriptor variables. Obviously, it is also impossible to "test a task" to see what scores it deserves on the various task descriptor variables. In such cases, the researcher may have to employ "task experts" (i.e., personnel from the real world of interest who are intimately familiar with the task domain in question) to not only help describe what tasks actually belong in a given task domain of interest, but also for estimating the relative frequencies with which those tasks occur in the task domain of interest. It should be recalled that each task included in a task domain of interest must be able to be "scored" on each of the task descriptor variables being used as part of the task domain description. The insistence on meeting this requirement is that almost any generic task (e.g., tracking, data entry, target recognition, etc.) has a large number of parameters on which that generic task can be varying and still be that kind of a task. A tracking task, for example, may be pursuit or compensatory, the target itself may be driven by a virtually infinite number of complex signals, the types of displays used to inform the tracker of the current situation may vary widely, the control devices furnished to the tracker for manipulating the acquisition of the target may also vary widely, and so forth. However, as long as one stays within the tracking task domain, the same variables should be used to describe any type of tracking task. The same observation can be made with regard to, say, a data entry task. That is, a large number of variables will be needed to precisely describe each possible kind of data entry task. But the variables needed for describing a tracking task are not the same as those needed to describe, say, data entry tasks. Indeed, many of the tracking task descriptor variables would be irrelevant for a data entry task and vice versa. Because of this, we can readily recognize that there are multiple task domains. A researcher may, of course, choose to include multiple task domains in a given study. This is certainly permissible, but each task domain included in the study must have its own set of task descriptor variables. And, within every task domain included, the "position" of each possible task in that domain must be estimated for each of its descriptor variables, and the frequency with which those tasks occur in the real world of interest should be specified. From such data, it is possible to calculate the interrelationships of the task descriptor variables. Volume 4 of this series gives examples of various task

domains that are applicable for a variety of jobs accomplished by Naval Flight Officers. An example of variables used to generically describe some of those task domains is also furnished in the document.

Similar determinations (of distributions and interrelationships of descriptor variables) should also be made for the environmental domain (or domains) under which the persons in the specified people domain must perform the tasks in the specified task domain(s). Many tasks done by humans are accomplished under conditions and in surroundings that would not differ significantly from those conditions and situations that typically exist in a laboratory setting. If this is the case, then the researcher should stipulate that, rather than going into an overly elaborate description of, say, what a typical office is like. However, many human tasks of interest must be conducted under environmental conditions which, unless they are properly simulated in the laboratory, would be quite different and those differences could significantly effect the performance of the people in the study.

Finally, if one is interested in generalizing the results to a particular real world of interest, a determination should be made for the interrelationships among the people, task, and environmental variables. That is to say, some "kinds" of people in the people domain of interest may do certain "kinds" of tasks in the task domain more frequently than do other kinds of people. Similarly, some kinds of tasks may be more prevalent under certain environmental conditions than others are. Again, estimates of this type of information may require "experts" from the field who have an intimate familiarity with what typically happens in the real world of interest.

The need for these aforementioned specifications of the domains of interest is that, as researchers, we would like to be able to "fill a large box" with the correct populations of people, task, and environment combinations so that, ultimately, we can randomly draw a sample from that box and collect our research data on the performance of those people doing those tasks in those environments. If we could do this, then, we know that performances obtained from that sample will yield unbiased estimates

of the performances for the entire joint populations of people, tasks, and environments of interest to us. The need for the detailed specifications of the distributions and interrelationships for the people, task, and environment descriptor variables is not for ultimately collecting data on each possible combination, but, rather, to be able to mathematically specify the populations of interest so that we may, at a later time, obtain an appropriate random sample from it.

5.3.1.1 The Number of Variables Needed to Describe a Domain

From the standpoint of being parsimonious, a researcher would prefer to adequately describe any given domain (people, tasks, or environments) with as few descriptor variables as possible. However, a researcher may not know what minimum sized set of variables will accomplish that goal for a given domain of interest. Usually, there will be a variety of ways to describe people, tasks, and environments. If a researcher initially uses more variables than needed to adequately describe a given domain, no harm will be done; some information will be overdetermined. If some variables contain information which overlaps that contained by other variables, this redundant information should be reflected in a relationship between the scores on those variables. While a researcher is not obligated to include every type of descriptive variable which could be used to describe a domain, it is certainly preferable to overdescribe the domain than to use so few variables that the domain is obviously incompletely or ambiguously described.

5.3.1.2 Format for the Domain Descriptor Variables

The important pieces of information needed about each domain descriptor variable, in addition to a verbal description of the variable, include the vector of applicable scores for that variable, the relative frequencies with which those scores are assumed to occur within the domain of interest, and the relationship of each variable with the other domain descriptors. The researcher also may wish to include the rationale for why each particular domain descriptor was chosen and how estimates of the frequencies and correlations with other descriptors were obtained (e.g., existing analyses, analyses based on random samples of available data, data collected by the investigator in order to be able to describe the

domain, "expert" estimates, etc.). If a descriptor variable does have a distribution which conforms (or is assumed to conform) to some theoretical distribution shape (e.g., normal, rectangular, or binomial), then inclusion of the mean and standard deviation for those variables contains sufficient information that the frequencies of various scores need not be given. A multinomial variable (i.e., one which has more than two possible scores) which does not conform to a well-known theoretical distribution can be sufficiently described by the vector of probabilities associated with each of the possible scores. A continuous multinomial variable can usually be adequately described with, say, ten to twelve discrete intervals.

5.3.2 Selecting the Situations to be Studied

Having derived the specified domains, the researcher now has specified all the variables which define the model of the real world which is to be the subject of the RSDV study. The researcher is now ready to select situations from that domain. This must be accomplished using random sampling if results of the study are to be generalized to the entire modeled real world. Use of the computer is especially helpful in doing this phase of the effort. Several possible methods will be discussed. Each method should lead to a sample of creatable situations which, taken together, will represent a random sample of the variation within all of the domains being studied. Volume 5 of this series contains some examples of how a computer can be used to randomly sample a fairly complex domain which requires many different variables to describe it.

5.3.2.1 Selecting the Sample of People to Study

In a human performance study, the actual people on whom data will be collected cannot be created, but should be acquired by either bringing into the laboratory a random sample of persons from the specified people domain, or by going into the field with a portable laboratory and collecting data on a selected sample from the field. This selected sample from the field must be carefully constructed so that its final composition closely matches the people domain specifications. If a given field site has more of a certain type of person than the domain as a whole, then relatively fewer numbers of that type of person should be selected.

Techniques for stratified random sampling to accomplish this purpose are well known and discussed in other sources.

5.3.2.2 Selecting the Sample of Tasks and Environments

At the time when the study is actually conducted, the selected tasks and environments to be presented to the selected persons being studied must be available. This means that whatever tasks are to be studied must be able to be simulated at the appropriate times. It must be remembered that the objective of the RSDV study is, in part, to determine how the various elements of the real world tasks and environments influence human behavior. Theoretically, this goal can be accomplished using either of two methods. The first method is referred to as the "real task" method. It involves simulating versions of the actual real world tasks in the laboratory. The second method is referred to as the "generic task" method and involves creating tasks which contain appropriate mixes of the elements of the real world tasks, but no single generic task may be exactly like any known real world task. The former method is sometimes easier to use because the actual real world tasks can be understood and simulated with relative ease. The latter method may be more difficult when it comes to trying to invent a generic task (for use in the laboratory) which has a combination of task elements in it which make it like real world tasks but not necessarily identical to any known real task.

There are excellent reasons for using either method. For example, when the real task method has been used, the validity of task performance in a laboratory study can be more easily determined by comparing performance on those same tasks in field situations. On the other hand, if the generic task method is used and performance on actual real world tasks in the field can be predicted by the results obtained from the laboratory study, then one should have greater confidence in the ability of the results to generalize and to predict task performance on new real world tasks which might occur in the future.

Selecting a random sample of so-called real tasks requires a somewhat different approach than that used for selecting a random sample

of generic tasks. If one uses the real task method, then the frequency with which each of the real tasks in that domain should be known. If we assume that there are K real tasks within the domain then the relative frequencies with which those tasks occur can be converted into probabilities such that the probabilities across all K tasks will sum to one. The tasks, themselves, can be randomly numbered from 1 to K and their respective probabilities can be entered into a vector having K entries. This vector can then be converted into a cumulative probability vector by adding the sum of all the preceding entries to each consecutive entry. This vector's final entry must, of course, be equal to one. To determine which of the K tasks to utilize on any given trial, the computer can generate a random deviate (a value having equal probability for all values between zero and one). The computer can, starting with the first entry in the cumulative probability vector, compare the generated random number with each entry in the cumulative probability vector. Whenever the generated random number is found to be equal to or less than the entry in the cell being compared, the process is stopped and that particular task is selected as the appropriate one. Following this simple procedure, which is extremely easy to program for a computer, will assure that the task selected is a random sample from the specified task domain. To select additional tasks, additional random numbers are generated and used as described above. The entire sample of tasks selected in this way must, by definition, be a random sample of the specified task domain, regardless of the number of tasks in the specified domain and regardless of the actual size of the sample drawn from that domain.

When the "real task" method is used, it is probably desirable to separately estimate the probable frequencies of the various possible environments in which each task must be done. Thus, each real world task in the domain can have a separate vector of probabilities of each possible environment. The computer, having determined a given task to use can now retrieve the appropriate environmental cumulative probability vector for that task, and, by generating another random deviate, can select the environment to pair with that selected task. This procedure also assures a random sample of the environment domain and allows any kind of suspected or known task-environment interaction to be properly represented in the

total sample of tasks and environments drawn. If it turns out that that there are no real interactions between tasks and environments (i.e., a particular environment would have the same probability of occurring for any selected task), then the same environmental cumulative probability vector would be applicable for all tasks, and the computer will not have to distinguish which task it has selected to be able to determine what environment is to be randomly paired with it. This is identical to saying that the environments are assumed to be independent of the tasks.

In the "real task" approach, the matrix of interrelationships among the task descriptor variables could be ignored and the tasks can be selected strictly on the basis of their probable frequencies. From our earlier discussion on the random sampling of items having multiple attributes, it should be obvious that if the selection of items from that domain is a random one, we should also have obtained a random sample of the values for each of the domain descriptor variables. The "generic task" approach can be thought of as the reverse of this process in that, for it, we first obtain a randomly selected value for each task domain descriptor variable and, then, use these variables' values to define a generic task. The procedure to accomplish this is very similar to that used in the "simulated data generation" technique discussed earlier. It starts by factor analyzing the interrelationships among the tasks and environments to obtain the independent common factors and to determine how much of the variance of each domain descriptor is being accounted for by those common factors. Additional "unique" independent factors are then created so that all of the variance of each variable is accounted for. For each factor (common and unique) a cumulative probability vector must then be created so that it can subsequently be used, in conjunction with a random deviate generated by the computer, to determine a standard score for each factor. To obtain each generic task description, the computer will generate a random deviate for each independent factor, convert those values into a standard scores for each factor, and multiply the standard scores by the respective factor loadings for each domain descriptor variable. This will result in a each generic task having an assigned value for each descriptor variable. The investigator can take these descriptions and create a generic task that matches those descriptions.

As mentioned earlier, every generic task selected in this fashion will fall somewhere within the task domain, but it may not correspond perfectly with any known real task. A similar statement can be made with regard to the generic environment derived in this manner. In essence, the generic-task (and generic-environment) method derives multiple attribute values for a task and environment that is to be simulated in the laboratory.

Since the generic tasks and environments are derived from randomly sampling the task and environment domains, as described by their respective domain descriptor variables, the obtained sample of generic tasks and environments specified by the above process will also represent a random sample of those domains. One possible approach is to have the computer determine a random sample of tasks and environments and then to present all of them to each member of the sample of persons being used. In this way, correlations across people for all sampled task and environment combinations can be computed. This correlation matrix can be factor analyzed to determine how many different independent factors are influencing performance in the modeled domains. The obtained factors can be treated as criterion variables to be predicted using multiple correlation where the people, task and environment descriptors are the potential predictors. This will permit a prediction equation to be developed that determines the best weights for those variables to predict the performances obtained in the study. These weights can also be used to predict how any of the actual real tasks would have been performed had they been used instead of generic tasks. Such scores could then be compared with actual performance of those same persons doing a sample of known real tasks.

6. SUMMARY

This document has presented the rationale and background for the Random Sampling of Domain Variance (RSDV) technique. The RSDV technique is seen as a powerful alternative to the Analysis of Variance (ANOVA) method, especially for the purposes of experimental design. The historical dissatisfaction with ANOVA results, and their traditional inability to generalize to real world problems was described. It was concluded that ANOVA, while being a valuable technique for investigating problems in which there are only a very few important variables which could effect human performance, was severely limited and restricted by its requirement for proportional cases in data cells and its requirement for forcing the experimental variables to be independent of each other. These requirements of ANOVA not only restrict its applicability to fairly simple problems, but they also frequently force the collection of data which may be misrepresentative of the real world situations of interest to the investigator.

The various decisions confronting investigators who are interested in conducting human performance studies were discussed. How those decisions can impact the results and conclusions reached by an investigator was also described. Decisions facing the investigator fall into three major categories: (a) what will be studied, (b) what performance measures will be recorded, and (c) what analysis methods will be used. One of the problems which became apparent during the development of the RSDV technique was that investigators sometimes have little specific understanding of the composition of the real world for which they desire to do research. Both the ANOVA technique (at least while using fixed and mixed effects models) and field study techniques do not require the investigator to specify the real world to which they would like to be able to generalize their results. Failure of investigators to specify the

people, task, and environment domains in human performance studies have resulted in their subsequent inability to recognize that data collected by them had little hope of being representative of the real world performances of humans for which they had professed an interest. The RSDV technique recognizes this drawback and makes domain specification a central part of conducting research studies of human performance.

The principles and a variety of applications of random sampling were described. The RSDV technique was shown to be a natural extension to experimental design randomization procedures and an obvious application of random sampling similar to that which underlies the theory of significance testing. The actual random sampling during an RSDV study is for the purpose of determining what task and environment combinations will be studied in laboratory situations so that the investigator can be assured of obtaining unbiased estimates of performance in the specified real world of interest. The RSDV technique is also seen to be a natural extension of mathematical modeling and simulation technology.

Procedures for how investigators might go about specifying the people, task, and environmental domains of interest for RSDV studies were described. The usage of computers to ensure appropriate random sampling from those specified domains was also discussed. Finally, how various multivariate procedures, including multiple correlation and factor analysis, can be used to analyze human performance data collected in RSDV studies was discussed.

REFERENCES

1. Bailey, D. E. *Probability and statistics: models for research*, New York: John Wiley & Sons, 1971.
2. Cochran, W. G. and Cox, G. M. *Experimental design*, New York: John Wiley & Sons, 1950 (Rev. 1957).
3. Darlington, R. B. Multiple regression in psychological research and practice. *Psychological Bulletin*, 1968, 69.
4. Dunnette, M. D. Fads, fashions, and folderol in psychology. *American Psychologist*, 1966, 21, 343-352.
5. Edwards, A. E. *Experimental design in psychological research*, New York: Holt, Rinehart and Winston, 1950 (Rev. 1960).
6. Fisher, R. A. Discussion on Dr. Wishart's paper. *J.R. Statistical Soc. Suppl.*, 1934, 1, 51-53.
7. Hays, W. L. *Statistics for the social sciences*, New York: Holt, Rinehart and Winston, 1972.
8. Hutchins, Jr., C. W. A new approach to the construction of a predictor battery. *The Ohio State University* (unpublished doctoral dissertation), 1970.
9. Lane, N. E. The influence of selected factors on shrinkage and overfit in multiple correlation. *The Ohio State University* (unpublished doctoral dissertation), 1971.
10. Nicholson, G. E. Prediction in future samples. In Okin et al (Eds). *Contributions to probability and statistics*. P.322-330. Palo Alto, CA: Stanford University Press, 1960.
11. Peters, C. C. and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw Hill, 1940.
12. Simon, C. W. Evaluation of basic and applied research: (1) pragmatic criteria. *Paper presented at 83rd annual APA convention*, Chicago, 1975.
13. Stone, E. F. and Hollenbeck, J. R. Some issues associated with the use of moderated regression. *Organizational Behavior & Human Performance*, 1984, 34, 195-213.
14. Wherry, Sr., R. J. A new formula for predicting the shrinkage of the multiple correlation coefficient. *Annals of Mathematical Statistics*, 1931, 2, 440-457.
15. Wherry, Sr., R. J. The Wherry-Doolittle test selection method. In Stead, Shartle et al (Eds). *Occupational Counseling Techniques*, Appendix 5. New York: American Book Company, 1940.
16. Wherry, Sr., R.J., Naylor, J. C., Wherry, Jr., R. J. and Fallis, R. F. Generating multiple samples of multivariate data with arbitrary population parameters. *Psychometrika*, 1965, 30, 303-313.
17. Wherry, Sr., R. J. *Contributions to correlational analysis*. Orlando, Florida: Academic Press, Inc., 1984.
18. Woodworth, R. S. and Schlosberg, H. *Experimental psychology*, New York: Henry Holt and Company, 1938 (Rev. 1954).